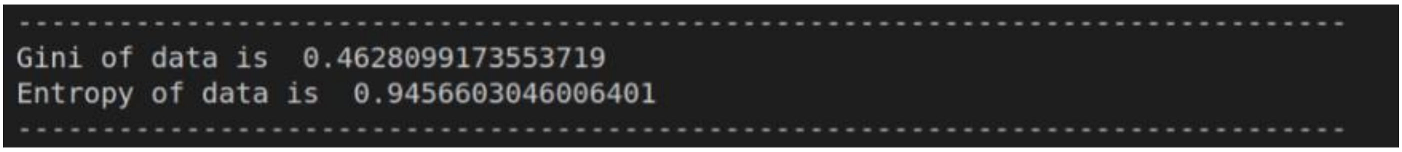
Decision tree algorithm (HW3)

Student: 張皓雲

Student Id: 310551031

● Part. 1, Coding (80%):

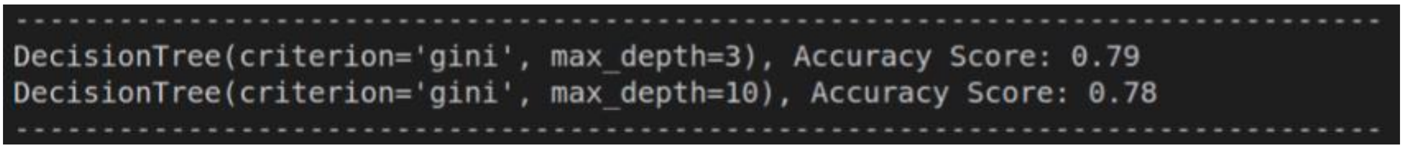
1. (5%) Gini Index or Entropy is often used for measuring the "best" splitting of the data. Please compute the Entropy and Gini Index of this array np.array([1,2,1,1,1,1,2,2,1,1,2]) by the formula below. (More details on page 5 of the hw3 slides, 1 and 2 represent class1 and class 2, respectively)



```
-----------------------------------------------------------------------
Gini of data is  0.4628099173553719
Entropy of data is  0.9456603046006401
-----------------------------------------------------------------------
```

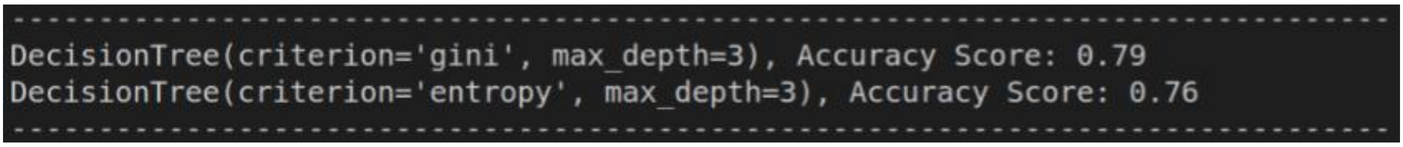
2. (10%) Implement the Decision Tree algorithm (CART, Classification and Regression Trees) and train the model by the given arguments, and print the accuracy score on the test data. You should implement two arguments for the Decision Tree algorithm,
   1) Criterion: The function to measure the quality of a split. Your model should support "gini" for the Gini impurity and "entropy" for the information gain.
   2) Max_depth: The maximum depth of the tree. If Max_depth=None, then nodes are expanded until all leaves are pure. Max_depth=1 equals split data once

   2.1. Using Criterion='gini', showing the accuracy score of test data by Max_depth=3 and Max_depth=10, respectively.



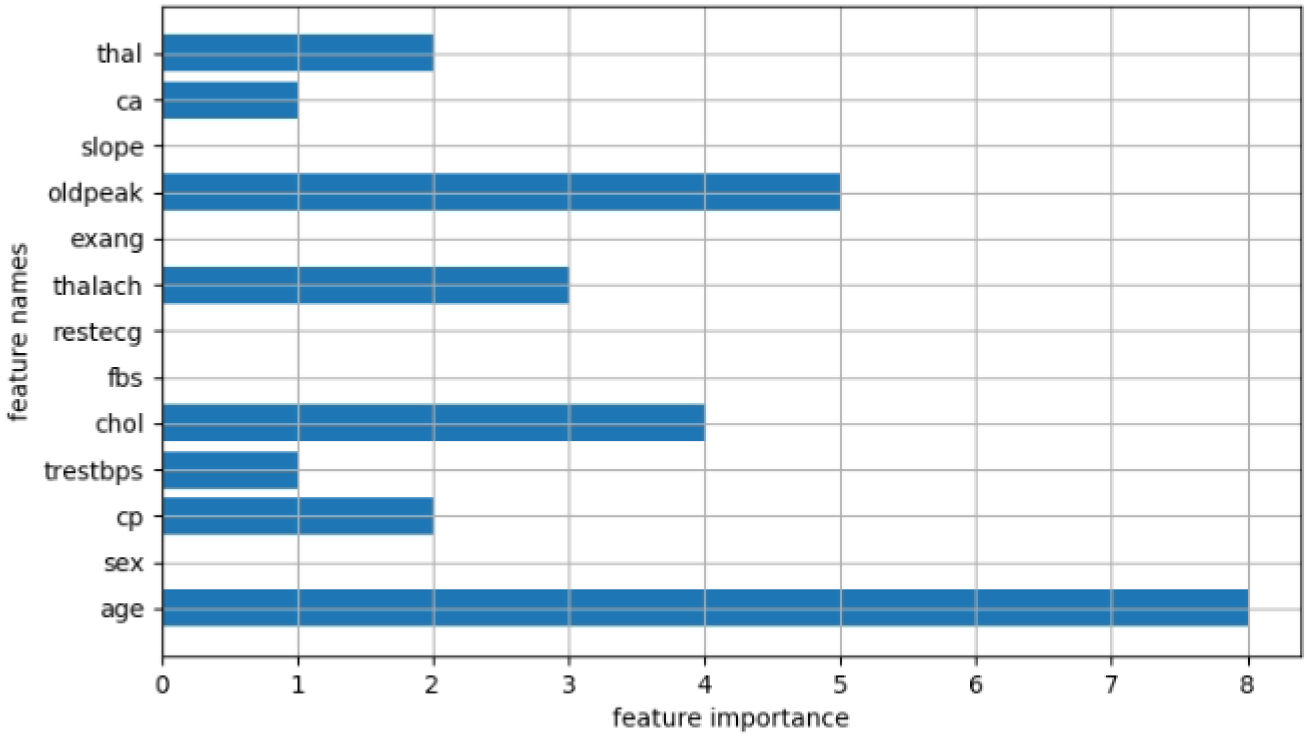
```
-----------------------------------------------------------------------
DecisionTree(criterion='gini', max_depth=3), Accuracy Score: 0.79
DecisionTree(criterion='gini', max_depth=10), Accuracy Score: 0.78
-----------------------------------------------------------------------
```

   2.2. Using Max_depth=3, showing the accuracy score of test data by Criterion='gini' and Criterion='entropy', respectively.

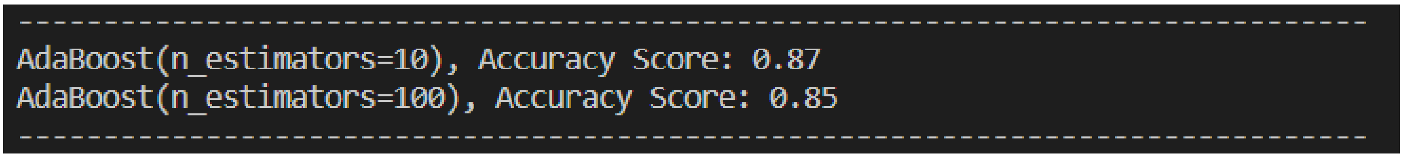


```
-----------------------------------------------------------------------
DecisionTree(criterion='gini', max_depth=3), Accuracy Score: 0.79
DecisionTree(criterion='entropy', max_depth=3), Accuracy Score: 0.76
-----------------------------------------------------------------------
```

3. (5%) Plot the feature importance of your Decision Tree model. You can use the model from Question 2.1, max_depth=10. (You can use simply counting to get the feature importance instead of the formula in the reference, more details on the sample code. **Matplotlib** is allowed to be used)



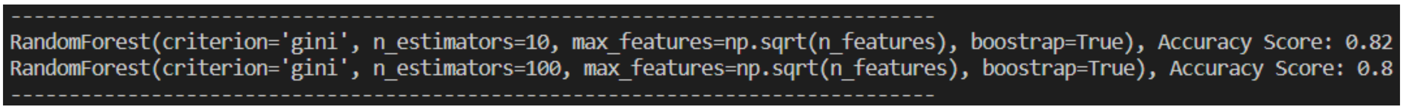
4. (15%) Implement the AdaBoost algorithm by using the CART you just implemented from question 2. You should implement one argument for the AdaBoost.
   1) N_estimators: The number of trees in the forest.
   4.1. Showing the accuracy score of test data by n_estimators=10 and n_estimators=100, respectively.



```
--------------------------------------------------------------------------
AdaBoost(n_estimators=10), Accuracy Score: 0.87
AdaBoost(n_estimators=100), Accuracy Score: 0.85
--------------------------------------------------------------------------
```

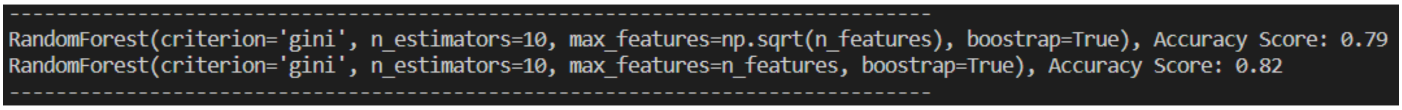
5. (15%) Implement the Random Forest algorithm by using the CART you just implemented from question 2. You should implement three arguments for the Random Forest.
   1) N_estimators: The number of trees in the forest.
   2) Max_features: The number of features to consider when looking for the best split
   3) Bootstrap: Whether bootstrap samples are used when building trees

   5.1. Using Criterion='gini', Max_depth=None, Max_features=sqrt(n_features), Bootstrap=True, showing the accuracy score of test data by n_estimators=10 and n_estimators=100, respectively.



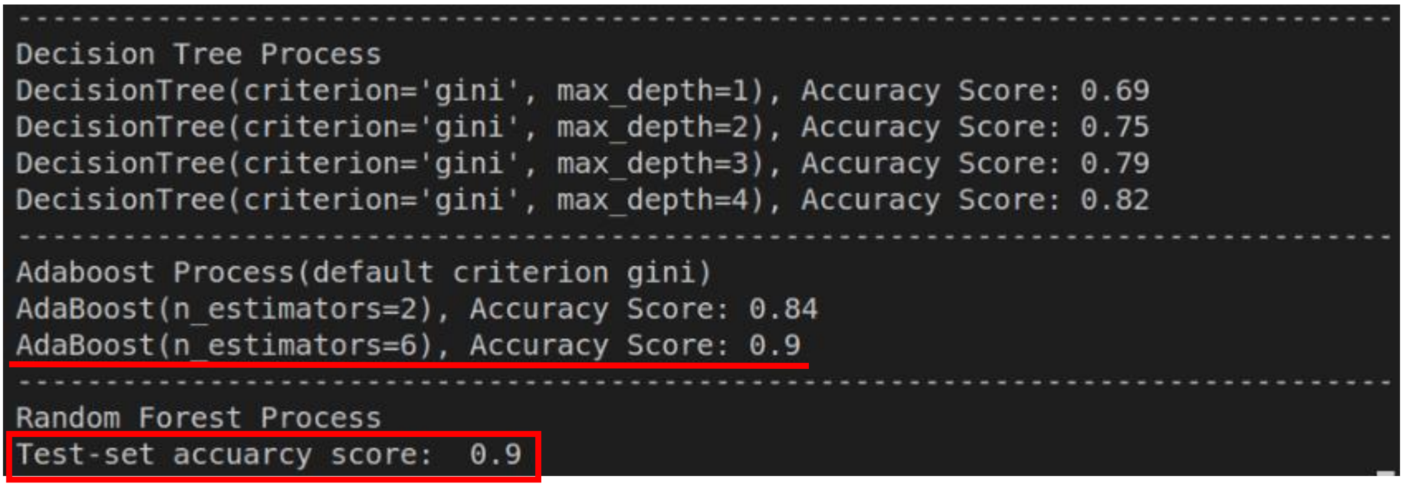
```
--------------------------------------------------------------------------
RandomForest(criterion='gini', n_estimators=10, max_features=np.sqrt(n_features), boostrap=True), Accuracy Score: 0.82
RandomForest(criterion='gini', n_estimators=100, max_features=np.sqrt(n_features), boostrap=True), Accuracy Score: 0.8
--------------------------------------------------------------------------
```

5.2. Using Criterion='gini', Max_depth=None, N_estimators=10, Bootstrap=True, showing the accuracy score of test data by Max_features=sqrt(n_features) and Max_features=n_features, respectively.



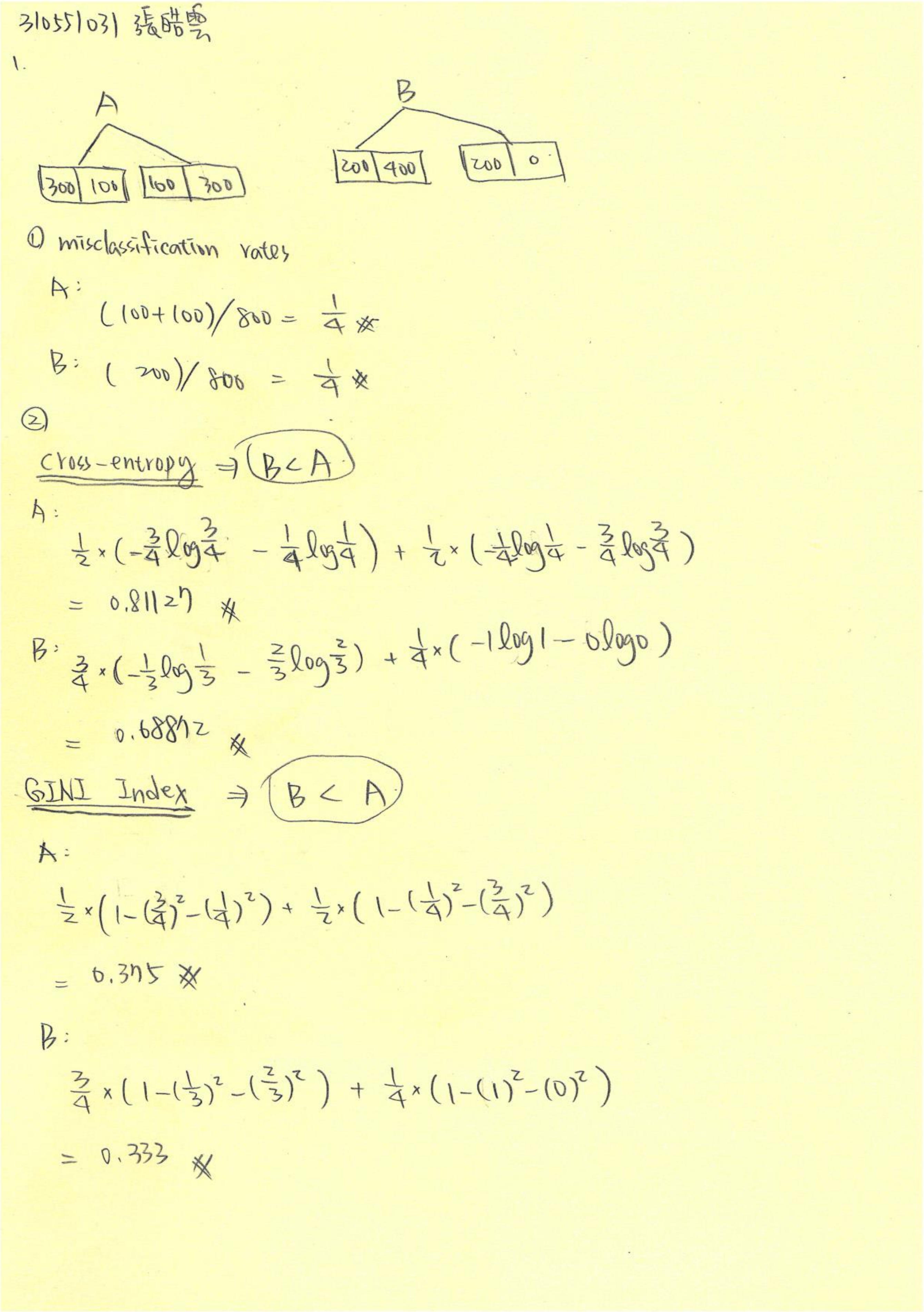
```
--------------------------------------------------------------------------------
RandomForest(criterion='gini', n_estimators=10, max_features=np.sqrt(n_features), boostrap=True), Accuracy Score: 0.79
RandomForest(criterion='gini', n_estimators=10, max_features=n_features, boostrap=True), Accuracy Score: 0.82
--------------------------------------------------------------------------------
```

6. (30%) Tune the hyperparameter, perform feature engineering or implement more powerful ensemble methods to get a higher accuracy score. Screenshot your tests score on the report. Please note that only the ensemble method can be used. The neural network method is not allowed.



```
--------------------------------------------------------------------------------
Decision Tree Process
DecisionTree(criterion='gini', max_depth=1), Accuracy Score: 0.69
DecisionTree(criterion='gini', max_depth=2), Accuracy Score: 0.75
DecisionTree(criterion='gini', max_depth=3), Accuracy Score: 0.79
DecisionTree(criterion='gini', max_depth=4), Accuracy Score: 0.82
--------------------------------------------------------------------------------
Adaboost Process(default criterion gini)
AdaBoost(n_estimators=2), Accuracy Score: 0.84
AdaBoost(n_estimators=6), Accuracy Score: 0.9
--------------------------------------------------------------------------------
Random Forest Process
Test-set accuarcy score:  0.9
```

● Part. 2, Questions (20%):



310551031 張晧雲

1.

A
300
100
100
300
B
200
400
200
0

① misclassification rates

A:

$(100+100)/800 = \frac{1}{4}$ ※

B: $(200)/800 = \frac{1}{4}$ ※

②

Cross-entropy $\Rightarrow$ $\boxed{B<A}$

A:

$\frac{1}{2}\times(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}) + \frac{1}{2}\times(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4})$

$= 0.81127$ ※

B: $\frac{3}{4}\times(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}) + \frac{1}{4}\times(-1\log 1 - 0\log 0)$

$= 0.68872$ ※

GINI Index $\Rightarrow$ $\boxed{B < A}$

A:

$\frac{1}{2}\times(1-(\frac{3}{4})^2-(\frac{1}{4})^2) + \frac{1}{2}\times(1-(\frac{1}{4})^2-(\frac{3}{4})^2)$

$= 0.375$ ※

B:

$\frac{3}{4}\times(1-(\frac{1}{3})^2-(\frac{2}{3})^2) + \frac{1}{4}\times(1-(1)^2-(0)^2)$

$= 0.333$ ※

2.

$$E_{x,t}[e^{-ty(x)}] = \sum_t \int e^{-ty(x)} P(t|x)\, P(x)\, dx$$

$$\frac{\partial E_{x,t}[e^{-ty(x)}]}{\partial x} = \sum_t e^{-ty(x)} P(t|x) P(x) = 0$$

$$e^{-y(x)} P(t=1|x) \cancel{P(x)} + e^{y(x)} P(t=-1|x) \cancel{P(x)} = 0$$

) 對y偏微.

$$-e^{-y(x)} P(t=1|x) + e^{y(x)} P(t=-1|x) = 0.$$

$$e^{2y(x)} = \frac{P(t=1|x)}{P(t=-1|x)}$$

$$y(x) = \frac{1}{2} \ln\left(\frac{P(t=1|x)}{P(t=-1|x)}\right) \#$$